



Pacini, D., & Windmeijer, F. (2016). Robust Inference for the Two-Sample 2SLS Estimator. *Economics Letters*, 146, 50-54.
<https://doi.org/10.1016/j.econlet.2016.06.033>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1016/j.econlet.2016.06.033](https://doi.org/10.1016/j.econlet.2016.06.033)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via Elsevier at <http://dx.doi.org/10.1016/j.econlet.2016.06.033>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>



Robust inference for the Two-Sample 2SLS estimator[☆]



David Pacini^{a,c}, Frank Windmeijer^{a,b,c,*}

^a Department of Economics, University of Bristol, UK

^b IEU, Bristol, UK

^c Cemmap, London, UK

HIGHLIGHTS

- We derive the variance of the TS2SLS estimator under heteroscedasticity.
- We propose a new robust variance estimator.
- We provide Stata code for the TS2SLS estimator and its robust variance estimator.
- We provide Stata code for an asymptotically equivalent nonlinear GMM estimator.

ARTICLE INFO

Article history:

Received 8 February 2016

Received in revised form

20 June 2016

Accepted 25 June 2016

Available online 19 July 2016

JEL classification:

C12

C13

C26

Keywords:

Linear model

Data combination

Instrumental variables

Robust inference

Nonlinear GMM

ABSTRACT

The Two-Sample Two-Stage Least Squares (TS2SLS) data combination estimator is a popular estimator for the parameters in linear models when not all variables are observed jointly in one single data set. Although the limiting normal distribution has been established, the asymptotic variance formula has only been stated explicitly in the literature for the case of conditional homoskedasticity. By using the fact that the TS2SLS estimator is a function of reduced form and first-stage OLS estimators, we derive the variance of the limiting normal distribution under conditional heteroskedasticity. A robust variance estimator is obtained, which generalises to cases with more general patterns of variable (non-)availability. Stata code and some Monte Carlo results are provided in an Appendix. Stata code for a nonlinear GMM estimator that is identical to the TS2SLS estimator in just identified models and asymptotically equivalent to the TS2SLS estimator in overidentified models is also provided there.

© 2016 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Two-Sample Two-Stage Least Squares (TS2SLS) estimator was introduced by Klevmarken (1982) and applies in cases where one wants to estimate the effects of possibly endogenous explanatory variables x on outcome y , but where y and x are not observed in the same data set. Instead, one has observations on outcomes y and instruments z in one sample (sample 1) and on x and z in another (sample 2). Related Two-Sample IV (TSIV)

estimators were proposed by Arellano and Meghir (1992) and Angrist and Krueger (1992). Furthermore, Angrist and Krueger (1995) proposed the TS2SLS estimator as a Split-Sample IV (SSIV) estimator. Inoue and Solon (2010) show that the TS2SLS estimator is more efficient than the TSIV estimator of Angrist and Krueger (1992). For further details, see Angrist and Pischke (2009) and the review of Ridder and Moffitt (2007).

This type of data combination estimation method is popular in economics. It is for example used in research on intergenerational mobility, as earnings of different generations are often not observed in the same data set, see the extensive list of references in Jerrim et al. (2014). A further recent application is van den Berg et al. (in press), who investigate the effect of early-life hunger on late-life health and use the two-sample IV approach to deal with imperfect recollection of conditions early in life. Pierce and Burgess (2013) propose the use of the TS2SLS estimator in epidemiology, in particular when estimating the

[☆] We would like to thank Helmut Farbmacher, Tom Palmer, Mark Schaffer, Jon Temple, Kate Tilling and the editor, Costas Meghir, for helpful comments. Windmeijer acknowledges funding by the Medical Research Council, grant no. MC_UU_12013/9.

* Corresponding author at: Department of Economics, University of Bristol, UK.
E-mail address: f.windmeijer@bristol.ac.uk (F. Windmeijer).

causal relationship between an exposure and an outcome using genetic factors as instrumental variables, so-called Mendelian randomisation, and where obtaining complete exposure data may be difficult due to high measurement costs.

Under certain assumptions, as stated below, the TS2SLS estimator is consistent and has a limiting normal distribution, see e.g. Klevmarken (1982) and Inoue and Solon (2010). Here we derive the limiting distribution of the TS2SLS estimator under general, unspecified, forms of conditional heteroskedasticity. As the TS2SLS estimator is a simple function of the reduced form parameters for y in sample 1, and the first-stage parameters for x in sample 2, its asymptotic variance is a function of the variances and covariances of these OLS estimators.

The variance of the limiting normal distribution of the TS2SLS estimator is given in (10) below and the formula for a robust estimator of the asymptotic variance is presented in (12). Neither of these have been derived and/or proposed in the literature before. The result in Inoue and Solon (2010) for the conditionally homoskedastic case is similar to our result for that case. They derive the limiting variance of the TS2SLS estimator from the optimal nonlinear GMM estimator. For overidentified models, these two estimators are not the same, but they have the same limiting distribution. Inoue and Solon (2010) did not derive the limiting robust variance for this GMM estimator, but did derive the limiting variance of the efficient two-step GMM estimator under general forms of conditional heteroskedasticity in Inoue and Solon (2005), which is also the approach presented in Arellano and Meghir (1992). Our derivation is different as we focus solely on the TS2SLS estimator as defined below in (5). For the conditional homoskedastic case, our variance estimator differs from the one proposed by Inoue and Solon (2010), as it uses the information from the two samples differently.

Applied researchers have constructed robust standard errors for the just-identified single endogenous regressor case by means of the delta method, see e.g. Dee and Evans (2003). Our result can be seen as a generalisation of this method to situations with multiple regressors and overidentification. Although we consider here a simple cross-sectional setup, other sampling designs can be accommodated and the result is straightforwardly extended to compute, for example, cluster-robust standard errors.

Our result also generalises to situations outside the standard TS2SLS setup. For example, it can accommodate a model with three explanatory variables where one endogenous variable is observed with the outcome variable in sample 1, but not in sample 2, one explanatory variable is only observed in sample 2 and one endogenous variable is observed in both samples 1 and 2. This is discussed in Section 5 below and we present Stata code for this example and for the standard TS2SLS setup in the Appendix (see Appendix A).

In the next section we present the model, assumptions and the TS2SLS estimator. In Section 3, we present our main results. Section 4 compares our results to those derived for nonlinear GMM. The Appendix also presents Stata code for the GMM estimator.

2. Model, assumptions and TS2SLS estimator

The structural linear model of interest is given by

$$y_i = x_i' \beta + \varepsilon_i, \quad (1)$$

but we cannot estimate this model as y_i and x_i are not jointly observed. Instead, we have two independent samples. In sample 1 we have observations on y and k_z exogenous instruments z . Sample 2 contains observations on the k_x explanatory variables x and z . Denoting by subscripts 1 and 2 whether the variables are observed in sample 1 or sample 2, in the first sample we observe $\{y_{1i}, z'_{1i}\}$ for $i = 1, \dots, n_1$, and in the second sample we observe $\{x'_{2j}, z'_{2j}\}$

for $j = 1, \dots, n_2$. Throughout we assume that $k_z \geq k_x$. Other explanatory variables that enter model (1), but that are observed in both samples and are exogenous, including the constant, have been partialled out.

The TS2SLS estimator is derived as follows. From the information in sample 1, we can estimate the reduced form model for y_{1i} , given by

$$y_{1i} = z'_{1i} \pi_{y1} + u_{1i}. \quad (2)$$

From sample 2, we can estimate the linear projections

$$x_{2j} = \Pi'_{x2} z_{2j} + v_{2j}, \quad (3)$$

with $\Pi_{x2} = E(z_{2j} z'_{2j})^{-1} E(z_{2j} x'_{2j})$, a $k_z \times k_x$ matrix of rank k_x by assumption. As (3) is a linear projection, it follows that $E(z_{2j} v'_{2j}) = 0$. Although the x_{1i} are not observed, the data generating process for y_{1i} is given by the structural model (1) and hence it and its reduced form are given by

$$\begin{aligned} y_{1i} &= x'_{1i} \beta + \varepsilon_{1i} = (z'_{1i} \Pi_{x1} + v'_{1i}) \beta + \varepsilon_{1i} \\ &= z'_{1i} \Pi_{x1} \beta + \varepsilon_{1i} + v'_{1i} \beta, \end{aligned} \quad (4)$$

with the linear projection parameters $\Pi_{x1} = E(z_{1i} z'_{1i})^{-1} E(z_{1i} x'_{1i})$. Again, $E(z_{1i} v'_{1i}) = 0$. From (2) and (4) it follows that $\pi_{y1} = \Pi_{x1} \beta$ and $u_{1i} = \varepsilon_{1i} + v'_{1i} \beta$. Clearly, knowledge of π_{y1} and Π_{x1} identifies the structural parameters β , and the standard 2SLS estimator in a sample with y_{1i} , x_{1i} and z_{1i} all observed combines the information contained in the OLS estimators for π_{y1} and Π_{x1} , denoted by $\hat{\pi}_{y1}$ and $\hat{\Pi}_{x1}$ as follows

$$\hat{\beta}_{2sls} = (\hat{\Pi}'_{x1} Z'_1 \hat{\Pi}_{x1})^{-1} \hat{\Pi}'_{x1} Z'_1 \hat{\pi}_{y1},$$

with Z_1 the $n_1 \times k_z$ matrix $[z'_{1i}]$.

As x_{1i} is not observed, we cannot estimate Π_{x1} , but we can estimate Π_{x2} using the second sample. Denoting the OLS estimator for Π_{x2} by $\hat{\Pi}_{x2}$, the Two-Sample 2SLS estimator is given by

$$\begin{aligned} \hat{\beta}_{ts2sls} &= (\hat{X}'_1 \hat{X}_1)^{-1} \hat{X}'_1 y_1 = (\hat{\Pi}'_{x2} Z'_2 \hat{\Pi}_{x2})^{-1} \hat{\Pi}'_{x2} Z'_2 y_1 \\ &= (\hat{\Pi}'_{x2} Z'_2 \hat{\Pi}_{x2})^{-1} \hat{\Pi}'_{x2} Z'_2 \hat{\pi}_{y1}. \end{aligned} \quad (5)$$

We make the following assumptions:

- A1: $\{y_{1i}, z'_{1i}\}_{i=1}^{n_1}$ and $\{x'_{2j}, z'_{2j}\}_{j=1}^{n_2}$ are i.i.d. random samples from the same population with finite fourth moments and are independent.
- A2: $E(z_{1i} z'_{1i}) = Q_{zz1}$; $E(z_{2j} z'_{2j}) = Q_{zz2}$. Q_{zz1} and Q_{zz2} are nonsingular.
- A3: $E(z_{1i} x'_{1i})$ and $E(z_{2j} x'_{2j})$ both have rank k_x .
- A4: $E(z_{1i} \varepsilon_{1i}) = 0$.
- A5: $E(u^2_{1i} z_{1i} z'_{1i}) = \Omega_{y1}$, a finite and positive definite matrix.
- A6: $E[(I_{k_x} \otimes z_{2j}) v_{2j} v'_{2j} (I_{k_x} \otimes z'_{2j})] = E(v_{2j} v'_{2j} \otimes z_{2j} z'_{2j}) = \Omega_{x2}$, a finite and positive definite matrix. I_{k_x} is the identity matrix of order k_x .
- A7: $\lim_{n_1 \rightarrow \infty, n_2 \rightarrow \infty} \frac{n_1}{n_2} = \alpha$ for some $\alpha > 0$.

Assumptions A1–A3 and A7 are standard data combination assumptions, see e.g. Inoue and Solon (2010). Assumptions A2 and A3, combined with A1, result in $E(z_{1i} z'_{1i}) = E(z_{2j} z'_{2j})$ and $E(z_{1i} x'_{1i}) = E(z_{2j} x'_{2j})$, and hence $\Pi_{x1} = \Pi_{x2}$. A1–A3 are clearly sufficient, but not necessary conditions for Π_{x1} to be equal to Π_{x2} . The condition $\Pi_{x1} = \Pi_{x2}$ itself is sufficient for consistency of $\hat{\beta}_{ts2sls}$, and necessary for the limiting normal distribution of $\sqrt{n_1}(\hat{\beta}_{ts2sls} - \beta)$ to have a mean of zero. In the derivations below we do not (need to) impose $Q_{zz1} = Q_{zz2}$. The resulting estimator of the variance of $\hat{\beta}_{ts2sls}$ is a simple function of the variances of $\hat{\pi}_{y1}$ and $\text{vec}(\hat{\Pi}_{x2})$, and this function is unambiguous about which information from which sample is being utilised.

Assumptions A5 and A6 explicitly allow for general forms of heteroskedasticity. The robust variance estimator for $\hat{\beta}_{ts2sls}$ is obtained incorporating robust variance estimators for $\hat{\pi}_{y1}$ and $\text{vec}(\hat{\Pi}_{x2})$. This was done by [Dee and Evans \(2003\)](#) using the delta method for the just identified single regressor case, i.e. $k_x = k_z = 1$. The result derived below can be seen as a generalisation of this to multiple regressors and overidentified settings.

3. Limiting distribution and variance estimator

The OLS estimators for π_{y1} and Π_{x2} are given by

$$\hat{\pi}_{y1} = (Z_1' Z_1)^{-1} Z_1' y_1$$

$$\hat{\Pi}_{x2} = (Z_2' Z_2)^{-1} Z_2' X_2,$$

with Z_1 the $n_1 \times k_z$ matrix $[z_{1i}']$; Z_2 the $n_2 \times k_z$ matrix $[z_{2j}']$; y_1 the n_1 vector (y_{1i}) and X_2 the $n_2 \times k_x$ matrix $[x_{2j}']$. Under Assumptions A1–A4 and A7 we obtain

$$\text{plim}(\hat{\pi}_{y1}) = E(z_{1i} z_{1i}')^{-1} E(z_{1i} x_{1i}') \beta$$

$$= \pi_{y1} = \Pi_{x1} \beta = \Pi_{x2} \beta;$$

$$\text{plim}(\hat{\Pi}_{x2}) = E(z_{2j} z_{2j}')^{-1} E(z_{2j} x_{2j}') = \Pi_{x2},$$

and hence the TS2SLS estimator is consistent as

$$\text{plim}(\hat{\beta}_{ts2sls}) = \text{plim} \left(\frac{1}{n_1} \hat{\Pi}_{x2}' Z_1' Z_1 \hat{\Pi}_{x2} \right)^{-1} \frac{1}{n_1} \hat{\Pi}_{x2}' Z_1' Z_1 \hat{\pi}_{y1}$$

$$= (\Pi_{x2}' Q_{zz1} \Pi_{x2})^{-1} \Pi_{x2}' Q_{zz1} \pi_{y1} = \beta. \quad (6)$$

Note that the probability limits obtained here and the limiting distributions derived below are for both $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$.

For the derivation of the asymptotic distribution of $\hat{\beta}_{ts2sls}$, denote $\pi_{x2} = \text{vec}(\Pi_{x2})$; $\hat{\pi}_{x2} = \text{vec}(\hat{\Pi}_{x2})$; $\theta = (\pi_{y1}' \ \pi_{x2}')'$ and $\hat{\theta} = (\hat{\pi}_{y1}' \ \hat{\pi}_{x2}')'$. Under Assumptions A1–A7

$$\sqrt{n_1}(\hat{\pi}_{y1} - \pi_{y1}) \xrightarrow{d} N(0, V_{\pi_{y1}}); \quad (7)$$

$$\sqrt{n_2}(\hat{\pi}_{x2} - \pi_{x2}) \xrightarrow{d} N(0, V_{\pi_{x2}}), \quad (8)$$

where

$$V_{\pi_{y1}} = Q_{zz1}^{-1} \Omega_{y1} Q_{zz1}^{-1};$$

$$V_{\pi_{x2}} = (I_{k_x} \otimes Q_{zz2}^{-1}) \Omega_{x2} (I_{k_x} \otimes Q_{zz2}^{-1}).$$

Hence

$$\sqrt{n_1}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_{\theta}), \quad (9)$$

with

$$V_{\theta} = \begin{bmatrix} V_{\pi_{y1}} & 0 \\ 0 & \alpha V_{\pi_{x2}} \end{bmatrix}.$$

From the limiting distribution of $\hat{\theta}$, the limiting distribution of $\hat{\beta}_{ts2sls}$ is readily obtained and we give a simple proof in the Appendix (see [Appendix A](#)). Our main result is:

Under Assumptions A1–A7, the limiting distribution of $\hat{\beta}_{ts2sls}$ is given by

$$\sqrt{n_1}(\hat{\beta}_{ts2sls} - \beta) \xrightarrow{d} N(0, V_{\beta});$$

$$V_{\beta} = C(V_{\pi_{y1}} + \alpha(\beta' \otimes I_{k_z})V_{\pi_{x2}}(\beta \otimes I_{k_z}))C'$$

$$= CV_{\pi_{y1}}C' + \alpha(\beta' \otimes C)V_{\pi_{x2}}(\beta \otimes C'), \quad (10)$$

where

$$C = (\Pi_{x2}' Q_{zz1} \Pi_{x2})^{-1} \Pi_{x2}' Q_{zz1}. \quad (11)$$

We can obtain an estimator for the asymptotic variance of $\hat{\beta}_{ts2sls}$ as follows. Let $\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1})$ and $\hat{V}_{\pi_{x2}}(\hat{\pi}_{x2})$ be estimators of the asymptotic variances of $\hat{\pi}_{y1}$ and $\hat{\pi}_{x2}$, in the sense that $\text{plim}(n_1 \hat{V}_{\pi_{y1}}(\hat{\pi}_{y1})) = V_{\pi_{y1}}$ and $\text{plim}(n_2 \hat{V}_{\pi_{x2}}(\hat{\pi}_{x2})) = V_{\pi_{x2}}$. Let \hat{C} be the matrix of least squares coefficients from the regressions of the columns of Z_1 on \hat{X}_1 . As $\text{plim}(\hat{C}) = \text{plim}((\hat{X}_1' \hat{X}_1)^{-1} \hat{X}_1' Z_1) = C$, an estimator of the asymptotic variance of $\hat{\beta}_{ts2sls}$ is given by

$$\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) = \hat{C} \hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) \hat{C}' + (\hat{\beta}_{ts2sls}' \otimes \hat{C})$$

$$\times \hat{V}_{\pi_{x2}}(\hat{\pi}_{x2}) (\hat{\beta}_{ts2sls} \otimes \hat{C}'), \quad (12)$$

as

$$n_1 \hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) = \hat{C} (n_1 \hat{V}_{\pi_{y1}}(\hat{\pi}_{y1})) \hat{C}' + \frac{n_1}{n_2} (\hat{\beta}_{ts2sls}' \otimes \hat{C})$$

$$\times (n_2 \hat{V}_{\pi_{x2}}(\hat{\pi}_{x2})) (\hat{\beta}_{ts2sls} \otimes \hat{C}')$$

$$\xrightarrow{p} V_{\beta}.$$

When the model is just identified, $k_z = k_x$, then $\hat{C} = \hat{\Pi}_{x2}^{-1}$. When furthermore $k_x = k_z = 1$, (12) reduces to the simple expression

$$\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) = (\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) + \hat{\beta}_{ts2sls}' \hat{V}_{\pi_{x2}}(\hat{\pi}_{x2})) / \hat{\pi}_{x2}^2,$$

with $\hat{\beta}_{ts2sls} = \frac{\hat{\pi}_{y1}}{\hat{\pi}_{x2}}$, which is identical to the expression obtained using the delta method as in [Dee and Evans \(2003\)](#).

Specifying $\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1})$ and $\hat{V}_{\pi_{x2}}(\hat{\pi}_{x2})$ in (12) as being robust to general forms of heteroskedasticity results in a robust variance estimator for $\hat{\beta}_{ts2sls}$. A small Monte Carlo exercise reported in the Appendix confirms that our asymptotic results reflect the behaviour of the TS2SLS estimator. Although we have here an i.i.d. cross-sectional setup, the results generalise to e.g. cluster-robust variances straightforwardly.

4. GMM

Assuming conditional homoskedasticity for both u_{1i} and v_{2j} such that

$$E(u_{1i}^2 | z_{1i}) = \sigma_u^2 \quad \text{and} \quad E(v_{2j} v_{2j}' | z_{2j}) = \Sigma_v,$$

we have that

$$V_{\pi_{y1}} = \sigma_u^2 Q_{zz1} \quad \text{and} \quad V_{\pi_{x2}} = \Sigma_v \otimes Q_{zz2}^{-1},$$

and hence

$$V_{\beta} = \sigma_u^2 (\Pi_{x2}' Q_{zz1} \Pi_{x2})^{-1} + \alpha \beta' \Sigma_v \beta C Q_{zz2}^{-1} C'.$$

The variance estimator (12) is then

$$\hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) = \hat{\sigma}_u^2 (\hat{X}_1' \hat{X}_1)^{-1}$$

$$+ \hat{\beta}_{ts2sls}' \hat{\Sigma}_v \hat{\beta}_{ts2sls} \hat{C} (Z_2' Z_2)^{-1} \hat{C}', \quad (13)$$

with $\hat{\sigma}_u^2 = (y_1 - Z_1 \hat{\pi}_{y1})' (y_1 - Z_1 \hat{\pi}_{y1}) / n_1$ and $\hat{\Sigma}_v = (X_2 - Z_2 \hat{\pi}_{x2})' (X_2 - Z_2 \hat{\pi}_{x2}) / n_2$.

[Inoue and Solon \(2010\)](#) derive V_{β} from the limiting distribution of the optimal GMM estimator using moment conditions

$$E[z_{1i} (y_{1i} - z_{1i}' \Pi_{x2} \beta)] = 0; \quad (14)$$

$$E[z_{2j} \otimes (x_{2j} - \Pi_{x2}' z_{2j})] = 0, \quad (15)$$

and weight matrix

$$\begin{bmatrix} \hat{V}_{\pi_{y1}}(\hat{\pi}_{y1}) & 0 \\ 0 & \hat{V}_{\pi_{x2}}(\hat{\pi}_{x2}) \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_u^2 (Z_1' Z_1)^{-1} & 0 \\ 0 & \hat{\Sigma}_v \otimes (Z_2' Z_2)^{-1} \end{bmatrix}.$$

Let $\psi = (\beta' \ \pi_{x2}')'$, then this GMM estimator is the same as the minimum distance estimator

$$\tilde{\psi} = \arg \min_{\beta, \pi_{x2}} \left(\begin{array}{c} \hat{\pi}_{y1} - \Pi_{x2}\beta \\ \hat{\pi}_{x2} - \pi_{x2} \end{array} \right)' \times \left[\begin{array}{cc} (V\hat{a}r(\hat{\pi}_{y1}))^{-1} & 0 \\ 0 & (V\hat{a}r(\hat{\pi}_{x2}))^{-1} \end{array} \right] \left(\begin{array}{c} \hat{\pi}_{y1} - \Pi_{x2}\beta \\ \hat{\pi}_{x2} - \pi_{x2} \end{array} \right).$$

Unless the model is just identified, $\tilde{\beta} \neq \hat{\beta}_{ts2s}$, but their limiting distributions are the same. This is a situation similar to that of the LIML and 2SLS estimators in the standard IV model. When the model is overidentified, the TS2SLS estimator itself cannot be obtained as a GMM estimator. The limiting variance of $\sqrt{n_1}(\tilde{\beta} - \beta)$ is obtained from the limiting variance of $\sqrt{n_1}(\tilde{\psi} - \psi)$. Inoue and Solon (2010) imposed $Q_{zz1} = Q_{zz2}$ and obtained the variance as

$$V_{\beta, IS} = (\sigma_u^2 + \alpha\beta' \Sigma_v \beta) (\Pi_{x2}' Q_{zz1} \Pi_{x2})^{-1}$$

and their variance estimator is given by

$$V\hat{a}r_{IS}(\hat{\beta}_{ts2s}) = \left(\hat{\sigma}_u^2 + \frac{n_1}{n_2} \hat{\beta}_{ts2s}' \hat{\Sigma}_v \hat{\beta}_{ts2s} \right) (\hat{X}_1' \hat{X}_1)^{-1},$$

where $\hat{\sigma}_u^2 = (y_1 - \hat{X}_1 \hat{\beta}_{ts2s})' (y_1 - \hat{X}_1 \hat{\beta}_{ts2s}) / n_1$. Apart from this difference in the estimation of σ_u^2 , the main difference is the imposition that $Q_{zz1} = Q_{zz2}$. Although this is justified asymptotically given the Assumptions A1–A3, the finite sample variance of $\hat{\pi}_{x2}$ in (12) is clearly more naturally estimated by $\hat{\Sigma}_v \otimes (Z_2' Z_2)^{-1}$ than by $\hat{\Sigma}_v \otimes (\frac{n_2}{n_1} Z_1' Z_1)^{-1}$. Also, for the example in footnotes 3 and 2 in Inoue and Solon (2010, 2005) respectively, when $E(z_{1i} x_{1i}') = cE(z_{2i} x_{2i}')$ and $E(z_{1i} z_{1i}') = cE(z_{2i} z_{2i}')$, with $c \neq 1$, then the TS2SLS estimator is consistent and asymptotically normally distributed but $n_1 V\hat{a}r_{IS}(\hat{\beta}_{ts2s})$ is no longer a consistent estimator of the variance of the limiting distribution, whereas $n_1 V\hat{a}r(\hat{\beta}_{ts2s})$ is.

Inoue and Solon (2010) did not derive the robust variance of $\tilde{\beta}$. Although this can be obtained from the robust variance of $\tilde{\psi}$, the matrix expressions involved are quite cumbersome. Arellano and Meghir (1992) similarly considered the robust variance of the GMM estimator $\tilde{\psi}$ but also did not derive a variance estimator for $\tilde{\beta}$ separately. One can of course simply obtain robust standard errors for $\tilde{\psi}$ and hence $\tilde{\beta}$ using GMM routines that can estimate the parameters using the nonlinear and linear moment conditions (14) and (15). These estimates are then obtained using iterative methods, and for just-identified models this produces the TS2SLS estimator with robust standard errors. For overidentified models, the efficient two-step GMM estimator for $\tilde{\psi}$ can then also be obtained together with a Hansen test for the validity of the moment conditions. We present Stata code for this GMM estimation procedure in the Appendix (see Appendix A).

5. Generalising the result

Although we derived the results in Section 3 for the standard TS2SLS estimator, the limiting distribution results (17) and (18) in the Appendix (see Appendix A) apply more generally. Indeed, the only aspect in V_θ that is particular to this specific two-sample setup is the zero covariance between $\hat{\pi}_{y1}$ and $\hat{\pi}_{x2}$, due to the samples being independent.

Consider as a generalisation a model with three explanatory variables x_1 , x_2 and x_3 . Using the same notational convention as before, in sample 1 we observe $\{y_{1i}, x_{11i}, x_{31i}, z_{1i}'\}_{i=1}^{n_1}$. In sample 2 we observe $\{x_{22j}, x_{32j}, z_{2j}'\}_{j=1}^{n_2}$. In this case, x_1 is only observed in sample 1, x_2 is only observed in sample 2, whereas x_3 is observed

in both samples. Let $Z = (Z_1' \ Z_2')'$ and $x_3 = (x_{31}' \ x_{32}')'$, then the reduced form and first-stage OLS estimators are given by

$$\hat{\pi}_{y1} = (Z_1' Z_1)^{-1} Z_1' y_1; \quad \hat{\pi}_{x11} = (Z_1' Z_1)^{-1} Z_1' x_{11}$$

$$\hat{\pi}_{x22} = (Z_2' Z_2)^{-1} Z_2' x_{22}; \quad \hat{\pi}_{x3} = (Z' Z)^{-1} Z' x_3.$$

Let $\hat{\Pi}_x = [\hat{\pi}_{x11} \ \hat{\pi}_{x22} \ \hat{\pi}_3]$, then the two-sample IV estimator is given by

$$\hat{\beta}_{2s} = (\hat{\Pi}_x' Z_1' Z_1 \hat{\Pi}_x)^{-1} \hat{\Pi}_x' Z_1' \hat{\pi}_{y1}.$$

We differentiate this estimator from the standard two-sample setup above and reserve the name $\hat{\beta}_{ts2s}$ for that particular setup. Under Assumptions A1–A7, the limiting distribution is as in (17), but as $\hat{\theta} = (\hat{\pi}_{y1}' \ \text{vec}(\hat{\Pi}_x)')'$, the variance V_θ differs from the standard setup as there is a different covariance structure. There are non-zero covariances between $\hat{\pi}_{y1}$ and $\hat{\pi}_{x11}$; $\hat{\pi}_{x11}$ and $\hat{\pi}_{x3}$; and $\hat{\pi}_{x11}$ and $\hat{\pi}_{x3}$, whereas the covariances between $\hat{\pi}_{y1}$ and $\hat{\pi}_{x22}$; and $\hat{\pi}_{x11}$ and $\hat{\pi}_{x22}$ are zero. From (18), an estimator for the asymptotic variance is given by

$$V\hat{a}r(\hat{\beta}_{2s}) = (\hat{\delta}' \otimes \hat{C}) V\hat{a}r(\hat{\theta}) (\hat{\delta} \otimes \hat{C}'), \quad (16)$$

where $\hat{\delta} = (1 \ -\hat{\beta}_{2s}')$ and $\hat{C} = (\hat{X}_1' \hat{X}_1)^{-1} \hat{X}_1' Z_1 = (\hat{\Pi}_x' Z_1' Z_1 \hat{\Pi}_x)^{-1} \hat{\Pi}_x' Z_1' Z_1$.

For the standard TS2SLS setup and the more general structures, one can obtain the robust variance estimates using standard routines. We give Stata code for two examples in the Appendix (see Appendix A). The structure of the algorithm for the general case is:

1. Estimate the reduced form and first-stage parameters by OLS, obtain the predicted values \hat{X}_1 and a robust variance estimate for $\hat{\theta} = (\hat{\pi}_{y1}' \ \hat{\pi}_{x2}')'$, the matrix $V\hat{a}r(\hat{\theta})$. In Stata, the latter can be obtained using the 'gmm' or the 'suest' routine.
2. Regress y_1 on \hat{X}_1 to obtain the TS2SLS estimator.
3. Regress the columns of Z_1 on \hat{X}_1 and collect the parameter estimates in the matrix \hat{C} .
4. Calculate $V\hat{a}r(\hat{\beta}_{2s})$ by the matrix expression in (16).
5. Some adjustments have to be made when parameters on exogenous variables and the constant are included in the estimation. These are detailed in the code in the Appendix (see Appendix A).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.econlet.2016.06.033>.

References

- Angrist, J.D., Krueger, A.B., 1992. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *J. Amer. Statist. Assoc.* 87, 328–336.
- Angrist, J.D., Krueger, A.B., 1995. Split-sample instrumental variables estimates of the return to schooling. *J. Bus. Econom. Statist.* 13, 225–235.
- Angrist, J.D., Pischke, J.-S., 2009. *Mostly Harmless Econometrics. An Empiricist's Companion*. Princeton University Press, Princeton.
- Arellano, M., Meghir, C., 1992. Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *Rev. Econom. Stud.* 59, 537–559.
- Dee, T.S., Evans, W.N., 2003. Teen drinking and educational attainment: Evidence from two-sample instrumental variables estimates. *J. Labor Econ.* 21, 178–209.
- Inoue, A., Solon, G., 2005. Two-sample instrumental variables estimators, NBER Technical Working Paper 311.
- Inoue, A., Solon, G., 2010. Two-sample instrumental variables estimators. *Rev. Econ. Stat.* 92, 557–561.
- Jerrim, J., Choi, A., Rodriguez, R.S., 2014. Two-Sample Two-Stage Least Squares (TS2SLS) estimates of earnings mobility: how consistent are they?, Working Paper No. 14-17, Institute of Education, University of London.

- Klevmarck, N.A., 1982. Missing variables and two-stage least squares estimation from more than one data set, Working Paper Series No. 62, Research Institute of Industrial Economics, Stockholm, Sweden.
- Pierce, B.L., Burgess, S., 2013. [Efficient design for Mendelian randomization studies: Subsample and 2-sample instrumental variables estimators](#). *Am. J. Epidemiol.* 178, 1177–1184.
- Ridder, G., Moffitt, R., 2007. [The econometrics of data combination](#). In: Heckman, J.J., Leamer, H.E. (Eds.), *Handbook of Econometrics* Vol. 6, Part B. pp. 5469–5547. (Chapter 75).
- van den Berg, G.J., Pinger, P.R., Schoch, J., 2015. Instrumental variable estimation of the causal effect of hunger early in life on health later in life. *Econom. J.* <http://dx.doi.org/10.1111/eoj.12250>, (in press).